Originally published in: Bański, Piotr/Barbaresi, Adrien/Clematide, Simon/Kupietz, Marc/Lüngen, Harald/Pisetta, Ines (Eds.): Proceedings of the LREC 2020 Workshop, Language Resources and Evaluation Conference, 11–16 May 2020, 8th Workshop on Challenges in the Management of Large Corpora (CMLC-8). - Paris: European Language Resources Association, 2020. Pp. 10-14.

Evaluating a Dependency Parser on DeReKo

Peter Fankhauser, Bich-Ngoc Do, Marc Kupietz
IDS Mannheim, Heidelberg University, IDS Mannheim
Germany, Germany, Germany
fankhauser@ids-mannheim.de, do@cl.uni-heidelberg.de, kupietz@ids-mannheim.de

Abstract

We evaluate a graph-based dependency parser on DeReKo, a large corpus of contemporary German. The dependency parser is trained on the German dataset from the SPMRL 2014 Shared Task which contains text from the news domain, whereas DeReKo also covers other domains including fiction, science, and technology. To avoid the need for costly manual annotation of the corpus, we use the parser's probability estimates for unlabeled and labeled attachment as main evaluation criterion. We show that these probability estimates are highly correlated with the actual attachment scores on a manually annotated test set. On this basis, we compare estimated parsing scores for the individual domains in DeReKo, and show that the scores decrease with increasing distance of a domain to the training corpus.

Keywords: Dependency Parsing, Large Corpora, Evaluation

1. Background and Aims

The Leibniz Institute for the German Language (IDS) has been building up the German Reference Corpus DeReKo (Kupietz et al., 2010) since its foundation in the mid-1960s and maintains it continuously. Since 2004, two new releases per year have been published. These are made available to the German linguistic community via the corpus analysis platforms COSMAS II (Bodmer, 2005) and KorAP (Bański et al., 2013), which allows the query and display of dependency annotations. DeReKo covers a broad spectrum of topics and text types (Kupietz et al., 2018). The latest release DeReKo 2020-I (Leibniz-Institut für Deutsche Sprache, 2020) contains 46.9 billion words. The number of registered users is about 45,000.

Linguistic Annotations in DeReKo DeReKo also features many linguistic annotation layers, including 4 different morphosyntactic annotations as well as one constituency and dependency annotation. The only dependency annotation is currently provided by the Maltparser (Nivre et al., 2006), however, based on a different dependency scheme. One of DeReKo's design principles is to distinguish between observations and interpretations. Accordingly (automatic) linguistic annotations are systematically handled as theory-dependent and potentially error-prone *interpretations*. DeReKo's approach to make them usable for linguistic applications is to offer several alternatives, ideally independent annotations (Belica et al., 2011) on all levels. With KorAP, users can then use the degree of agreement between alternative annotations to get an idea of the accuracy they can expect for specific queries and query combinations. By using disjunctive or conjunctive queries on annotation alternatives, users can, in addition, try to maximise recall or precision, respectively (Kupietz et al., 2017). With this approach, the direct comparison of the average accuracy of two annotation tools or models does not play a decisive role, since normally one would add both variants anyway. However, since DeReKo is first of all very large and secondly permanently extended and improved, it is a prerequisite that an annotation tool is sufficiently performant to be applicable to DeReKo or to additional corpus text within reasonable time. This is not always the case, especially with syntactic annotations.

Given this background, the evaluation criteria for dependency annotations might differ from those in other applications. Important factors are above all: 1) sufficient performance and stability of the annotation tool; 2) independence from existing annotations; 3) at least selective improvements over existing annotations 4) Adaptability to domains outside the training data

2. Parser and Corpora

Parser The evaluated parser is a re-implementation of the graph-based dependency parser from Dozat and Manning (2017). The parser employs several layers of bidirectional Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) units to encode the words in a sentence. These representations are then used to train two biaffine classifiers, one to predict the head of a word and the other to predict the dependency label between two words. At prediction time, the dependency head and label for each word is selected as the word and label with the highest estimates given by the classifiers. The parser is available on Github (Do, 2019).

Training data We train the parser on the German dataset of the SPMRL 2014 Shared Task (Seddah et al., 2014) with the hyperparameters recommended by the authors. The dataset contains 40,000 sentences (760000 tokens) in the training set and 5,000 sentences (81700, 97000 tokens) for both development and testing. We use the predicted POS tags provided by the shared task organizers. For some evaluations we also use external word embeddings (see Section 3.) trained on DeReKo.

Evaluation data As evaluation data we use a sample of release 2019-I (Leibniz-Institut für Deutsche Sprache, 2019) of the German Reference Corpus DeReKo with 3670 Mio tokens from 11 domains. For a breakdown see Table 3. The corpus has been tokenized and part-of-speech tagged by the treetagger (Schmid, 1994). Parsing the corpus on a TESLA P4 GPU (8 GB) takes about 100 hours. For comparison, parsing with Malt 1.9.2 (liblinear) takes 34 wall-clock hours (38 CPU-hours) on the same machine equipped with enough

Proceedings of he 8th Workshop on Challenges in the Management of Large Corpora, pages 10–14
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

RAM and Intel Xeon Gold 6148 CPUs (at 2.40 GHz), when the corpus is processed sequentially.

This means that parsing with the malt parser is much more performant, especially since it can be distributed more easily to several existing computers and cores. On the other hand, parsing with the biaffine LSTM parser is at least sufficiently performant in the case of DeReKo. By using an additional GPU, DeReKo could be parsed within less than 4 weeks.

3. Overall Accuracy

As basic measures for parsing accuracy we use unlabeled and labeled attachment scores, UAS and LAS. UAS gives the percentage of dependency relations with the correct head and dependent, and LAS the percentage of correctly attached and labelled dependencies. In addition, we also look at the attachment estimates given by the two biaffine classifiers of the parser (see Equations 2 and 3 in Dozat and Manning (2017)). The estimates for the head of a dependency (unlabeled attachment estimate, UAE) and for its label (independent labeled attachment estimate, ILAE) are independent. Thus we calculate the labeled attachment estimate LAE as the product of UAE and ILAE.

External Word Embeddings Table 1 compares the attachment scores and estimates for different embeddings on the test set. For SPMRL embeddings we have experimented with embedding dimensions 100 and 200, for DEREKO embeddings we have used 200 dimensions throughout. The internal SPMRL embeddings are trained as part of the parser training process, the DEREKO embeddings have been trained using the structured skip gram approach introduced in (Ling et al., 2015) on the complete DeReKo-2017-I corpus (Institut für Deutsche Sprache, 2017) consisting of over 30 billion tokens. DEREKO1 uses the embeddings for the most frequent 100.000 words, DEREKO2 and DEREKO5 the most frequent 200.000 and most frequent 500.000 words respectively. The best overall scores are achieved with DEREKO2 leading to an improvement of about 0.5% in UAS and 0.8% in LAS w.r.t. the baseline of SPMRL without external embeddings. Taking into account a larger vocabulary (DEREKO5) does not improve the scores, nor does concatenating the internal embeddings of the parser with the DEREKO embeddings DRK2+SPMRL.

Scores vs. Estimates Comparing the scores with the parsers' estimates along varying embeddings also shows that they are highly correlated with the spearman rank correlation coefficient $\rho = 0.89$ between UAS and UAE, and $\rho = 0.94$ between LAS and LAE.

embeddings	dim	UAS	LAS	UAE	LAE
SPMRL	100	93.99	92.33	95.84	94.11
SPMRL	200	94.15	92.59	96.23	94.66
DEREKO1	200	94.30	93.00	97.08	95.90
DEREKO2	200	**94.51**	**93.16**	**97.10**	**95.94**
DEREKO5	200	93.98	92.50	95.88	94.40
DRK2+SPMRL	200	94.02	92.58	96.97	95.79

Table 1: Attachment scores and estimates for different word embeddings

All further evaluations use the model with the best scores DEREKO2.

Figure 1 plots the attachment scores against the attachment estimates between 75% and 100% in bins of 1%, i.e., the value at 99% estimate is the average score of all attachments with an estimate between 99% and 100%, and so on, and estimates smaller than 75% are bundled together with an average score of about 50%. Blue boxes stand for UAS and red circles for the LAS. Also from this perspective, the estimates strongly correlate with the scores. However, the estimates are typically overly confident. For the about 70% (63%) of attachments with an unlabeled (labeled) estimate $\geq$ 99% we get 99.79% UAS and 99.84% LAS. For the about 15% attachments with estimates between 98% and 99%, UAS and LAS are at about 96%. For lower estimates the difference between estimate and actual score increases. Nevertheless, the estimates predict the actual scores rather well, with Spearman's $\rho = 0.94$ for UAE vs. UAS, and $\rho = 0.99$ for LAE vs. LAS.

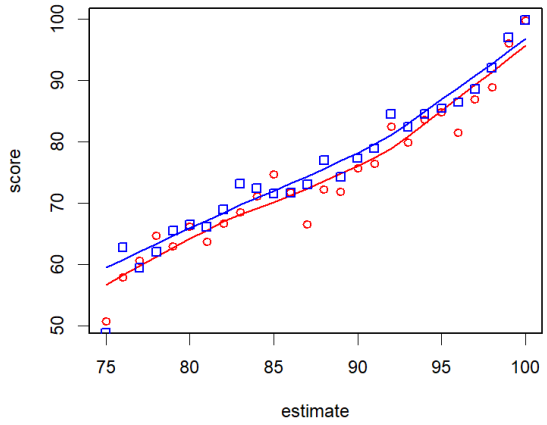

Figure 1: Attachment Estimates vs. Scores

4. Breakdown by Dependency Label

Table 2 breaks down scores and estimates by dependency label[1]. PROB gives the relative frequency of a dependency label in percent, UERR gives the percentage of overall error for unlabeled attachment, LERR the percentage for labeled attachment, REC the recall and PREC the precision for labeled attachment only, not taking into account the correctness of head and dependent.

In terms of individual scores, relatively rare dependencies such as Parataxes or Appositions perform worst. However, the frequency PROB of dependencies does not seem to have a strong influence on score, $\rho = -0.05$ for UAS vs. PROB, and $\rho = 0.42$ for LAS vs. PROB.

In terms of contribution to the overall error, Modifier (MO), Modifier of NP to the right (MNR), and Punctuation (X..) account for more than 50%. MO is often mislabelled as MNR or Object Preposition (OP) and vice versa, which typically also assigns the head incorrectly, as evident by the

[1]The SPMRL 2014 Shared Task for German uses the dependency scheme adopted by Seeker and Kuhn (2012)

rather low UAS of 88%. Punctuation is virtually never confused with other labels, its score of 91% is almost exclusively due to incorrect head or dependent attachments.

In terms of recall, rare dependencies such as Vocative (VO), Reported Speech (RS), and Object Genitive (OG) stand out, e.g. only 1 out of 15 occurrences of Vocative is correctly labeled, and less than half of RS and OG. Also, rare dependencies tend to depict low precision.

Comparing scores with estimates broken down by dependency label again reveals a rather strong correlation of $\rho = 0.89$ for unlabeled and $\rho = 0.75$ for labeled attachments.

5. Domain Dependence

Having established attachment estimates as a fairly reliable predictor for attachment scores, we can derive estimates for DEREKO for which we do not have any test data.

Table 3 breaks down estimates by domain, sorted by UAE. It can be seen that domains that are close to the news domain, for which the parser has been trained, such as politics, finance, and health achieve the best overall estimates. In contrast, domains, such as fiction, culture, and sports depict significantly lower estimates.

domain	UAE	LAE	JS_dep	JS_pos	Mio tokens
politics	95.85	95.14	0.13	0.24	820
finance	95.75	95.05	0.20	0.54	219
health	95.74	94.98	0.18	0.47	66
science	95.56	94.81	0.18	0.44	140
society	95.34	94.66	0.40	0.68	841
technology	95.18	94.50	0.15	0.45	196
leisure	95.15	94.43	0.24	0.32	469
nature	95.04	94.33	0.57	0.87	0.17
culture	94.52	93.79	0.41	0.31	453
sports	94.12	93.59	0.60	0.77	464
fiction	92.66	92.16	2.03	2.47	0.43

Table 3: Attachment estimates by domain

One way to measure the distance between domains w.r.t. to dependencies is to compare their distributions over dependency labels. JS_dep gives the Jensen-Shannon Divergence ($*100$) between the dependency distributions of the individual domains in DEREKO and the SPMRL training corpus. The closest is politics, and the most distant is fiction. Indeed, we can observe a strong negative correlation between UAE and JS_dep of -0.92 (Pearson) and LAE and JS_dep of -0.84. These findings are corroborated by the likewise fairly strong negative correlations between attachment estimates and JS_pos the JS divergence measured on the part-of-speech distributions; -0.48 for UAE and -0.84 for LAE.

6. Summary

We have presented an evaluation of a graph-based dependency parser on a large corpus of contemporary German for which no manually labelled test set is available. To this end, we have analyzed the correlation between actual attachment scores measured on the SPMRL test set with the parser's attachment estimates, and shown that they are highly correlated along variations in pretrained word embeddings (Table 1), as well as along the different kinds of dependencies (Table 2). On this basis, we have shown that the parser's attachment estimates are consistently domain dependent, with estimates varying up to 3% depending on distance of the domain to the training set. This suggests that it may be fruitful to experiment with domain adaptation techniques such as (Yu et al., 2015) in order to improve scores. For future work, we plan to systematically compare scores and estimates with the Malt parser. Depending on the results, we plan to apply the parser to the entire DeReKo in one of the upcoming releases and make the new dependency annotation layer available to German linguistics for research and analysis via KorAP.

7. Bibliographical References

Bański, P., Bingel, J., Diewald, N., Frick, E., Hanl, M., Kupietz, M., Pezik, P., Schnober, C., and Witt, A. (2013). KorAP: the new corpus analysis platform at IDS Mannheim. In Vetulani, Z. and Uszkoreit, H., editors, *Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 6th Language and Technology Conference*, Poznań. Fundacja Uniwersytetu im. A. Mickiewicza.

Belica, C., Kupietz, M., Lüngen, H., and Witt, A. (2011). The morphosyntactic annotation of DEREKO: Interpretation, opportunities and pitfalls. In Konopka, M., Kubczak, J., Mair, C., Šticha, F., and Wassner, U., editors, *Selected contributions from the conference Grammar and Corpora 2009*, pages 451–471, Tübingen. Gunter Narr Verlag.

Bodmer, F. (2005). COSMAS II. Recherchieren in den Korpora des IDS. *Sprachreport*, 3/2005:2–5.

Do, B.-N. (2019). Theano biaffine dependency parser. https://github.com/bichngocdo/theano-biaffine-parser. Accessed 2020-02-20.

Dozat, T. and Manning, C. D. (2017). Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Kupietz, M., Belica, C., Keibel, H., and Witt, A. (2010). The German Reference Corpus DEREKO: A Primordial Sample for Linguistic Research. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 1848–1854, Valletta/Paris. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf.

Kupietz, M., Diewald, N., Hanl, M., and Margaretha, E. (2017). Möglichkeiten der Erforschung grammatischer Variation mithilfe von KorAP. In Konopka, M. and Wöllstein, A., editors, *Grammatische Variation. Empirische Zugänge und theoretische Modellierung*, pages 319–329.

lb	meaning	UAS	LAS	UAE	LAE	Prob	Uerr	Lerr	Rec	Prec
AC	Adpositional Case Marker	95.38	95.38	99.16	99.10	0.14	0.12	0.09	99.21	96.92
ADC	Adjective Component	100.00	75.00	100.00	100.00	0.00	0.00	0.00	75.00	75.00
AG	Attribute Genitive	97.96	97.25	97.85	97.08	2.45	0.91	0.99	98.62	98.18
AMS	Measure Argument of Adjective	95.12	89.02	95.36	92.52	0.09	0.08	0.14	97.33	89.02
APP	Apposition	78.64	67.73	85.92	74.18	0.48	1.87	2.27	71.58	75.00
AVC	Adverbial Phrase Component	66.67	66.67	65.50	64.91	0.00	0.00	0.00	50.00	66.67
CC	Comparative Complement	84.74	84.34	89.17	87.25	0.27	0.75	0.62	93.28	89.16
CD	Coordinating Conjunction	93.08	92.99	95.33	95.24	2.43	3.06	2.49	99.82	99.42
CJ	Conjunct	91.10	89.64	94.33	92.48	3.72	6.03	5.64	91.56	92.09
CM	Comparative Conjunction	97.97	97.97	97.65	97.65	0.32	0.12	0.10	99.33	100.00
CP	Complementizer	99.24	99.24	99.52	99.48	0.86	0.12	0.10	100.00	100.00
CVC	Collocational Verb Construction	98.70	77.92	99.23	86.23	0.08	0.02	0.26	84.51	77.92
DA	Dative	94.95	90.09	95.68	88.33	0.58	0.53	0.84	87.50	92.90
DM	Discourse Marker	80.00	73.33	88.50	84.04	0.02	0.07	0.08	66.67	80.00
EP	Expletive	100.00	88.60	99.42	87.96	0.21	0.00	0.35	91.94	88.60
JU	Junctor	89.95	89.95	96.12	95.64	0.24	0.44	0.35	95.18	99.09
MNR	Modifier of Np to the right	78.77	75.20	84.97	81.05	2.84	10.98	10.30	84.03	82.25
MO	Modifier	88.46	86.65	90.60	87.81	13.01	27.35	25.40	93.73	94.75
NG	Negation	82.43	82.43	89.44	89.43	0.56	1.79	1.44	99.81	99.03
NK	Noun Kernerl Modifier	99.29	99.14	99.53	99.27	30.32	3.92	3.81	99.46	99.48
NMC	Numerical Component	99.69	98.75	99.61	99.15	0.35	0.02	0.06	98.75	98.75
OA	Object Accusative	97.00	92.74	97.01	92.56	3.55	1.94	3.77	96.11	93.69
OC	Object Clausal	97.83	95.11	97.97	95.80	4.00	1.58	2.86	96.71	95.93
OG	Object Genitive	100.00	71.43	90.93	76.07	0.02	0.00	0.08	47.62	71.43
OP	Object Preposition	95.85	72.89	96.21	75.99	0.73	0.55	2.89	76.47	73.19
PAR	Parataxis	62.20	50.40	76.51	62.22	0.41	2.82	2.97	56.64	65.15
PD	Predicative	98.05	90.33	98.24	90.21	1.11	0.39	1.57	88.90	90.72
PG	Pseudo Genitive	94.13	89.87	94.51	87.18	0.41	0.44	0.61	89.43	92.53
PH	Placeholder	100.00	86.21	99.70	73.44	0.03	0.00	0.06	83.33	86.21
PM	Morphological Particle	100.00	100.00	100.00	100.00	0.47	0.00	0.00	99.77	100.00
PNC	Proper Noun Component	96.16	95.04	97.73	96.44	1.36	0.95	0.99	95.91	95.60
RC	Relative Clause	83.48	82.84	88.61	88.08	0.84	2.53	2.11	98.82	97.55
RE	Repeated Element	87.86	87.50	93.34	91.54	0.30	0.66	0.55	91.79	87.86
RS	Reported Speech	85.19	55.56	88.35	73.58	0.03	0.08	0.20	42.86	55.56
RT	Root	94.97	94.97	98.66	98.29	5.94	5.44	4.37	97.35	94.97
SB	Subject	98.53	96.99	98.72	96.58	7.18	1.92	3.16	96.79	97.20
SBP	Subject Passivized	92.66	81.36	95.09	84.73	0.19	0.25	0.52	92.31	81.36
SVP	Separable Verb Prefix	99.40	99.00	99.36	99.23	0.54	0.06	0.08	99.80	99.60
UC	(Idiosyncratic) unit component	74.19	69.89	87.14	85.83	0.10	0.47	0.44	84.44	81.72
VO	Vocative	100.00	100.00	98.33	66.42	0.00	0.00	0.00	6.67	100.00
X..	Other (Punctuation)	91.36	91.36	95.04	95.04	13.80	21.72	17.44	99.30	99.76

Table 2: Scores and Estimates by Dependency Label

De Gruyter, Berlin. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-59681.

Kupietz, M., Lüngen, H., Kamocki, P., and Witt, A. (2018). The German Reference Corpus DeReKo: New Developments – New Opportunities. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18)*, pages 4353–4360, Miyazaki/Paris. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2018/pdf/737.pdf.

Ling, W., Dyer, C., Black, A., and Trancoso, I. (2015). Two/too simple adaptations of word2vec for syntax problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015)*, pages 1299—1304, Denver, CO.

Nivre, J., Hall, J., and Nilsson, J. (2006). Maltparser: A data-driven parser-generator for dependency parsing. In *LREC*, volume 6, pages 2216–2219.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf.

Seddah, D., Kübler, S., and Tsarfaty, R. (2014). Introducing the SPMRL 2014 shared task on parsing morphologically-rich languages. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pages 103–109, Dublin, Ireland. Dublin City University.

Seeker, W. and Kuhn, J. (2012). Making ellipses explicit in dependency conversion for a German treebank. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3132–3139, Istanbul, Turkey. European Language Resources Association (ELRA).

Yu, J., Elkaref, M., and Bohnet, B. (2015). Domain adaptation for dependency parsing via self-training. In *Proceedings of the 14th International Conference on Parsing Technologies*, pages 1–10, Bilbao, Spain. Association for Computational Linguistics.

8. Language Resource References

Institut für Deutsche Sprache (2017). German Reference Corpus DeReKo-2017-I. PID: http://hdl.handle.net/10932/00-0373-23CD-C58F-FF01-3.

Leibniz-Institut für Deutsche Sprache (2019). German Reference Corpus DeReKo-2019-I. PID: http://hdl.handle.net/10932/00-04BB-AF28-4A4A-2801-5.

Leibniz-Institut für Deutsche Sprache (2020). German Reference Corpus DeReKo-2020-I. PID: http://hdl.handle.net/10932/00-04B6-B898-AD1A-8101-4.